2017 The State of Data Science & Machine Learning

This year, for the first time, we conducted an industry-wide survey to establish a comprehensive view of the state of data science and machine learning. We received over **16,000 responses** and learned a ton about who is working with data, what's happening at the cutting edge of machine learning across industries, and how new data scientists can best break into the field. The below report shares some of our key findings and includes interactive visualizations so you can easily cut the data to find out exactly what you want to know. Here are some sample takeaways:

1. While Python may be the most commonly used tool overall, more Statisticians report using R.
2. On average, data scientists are around 30 years old, but this value varies between countries. For instance, the average respondent from India was about 9 years younger than the average respondent from Australia.
3. The highest percentage of our respondents obtained a Master's degree, but those in the highest salary ranges ($150K+) are slightly more likely to have a doctoral degree.

We've shared the full, anonymized dataset on Kaggle for you to download and explore. To participate in the conversation, share your analyses and code alongside the data so together we can continue advancing the state of data science and machine learning. You can even win cash prizes for your work. Who is Kaggle?

Download the survey data

Get the R kernel for this report

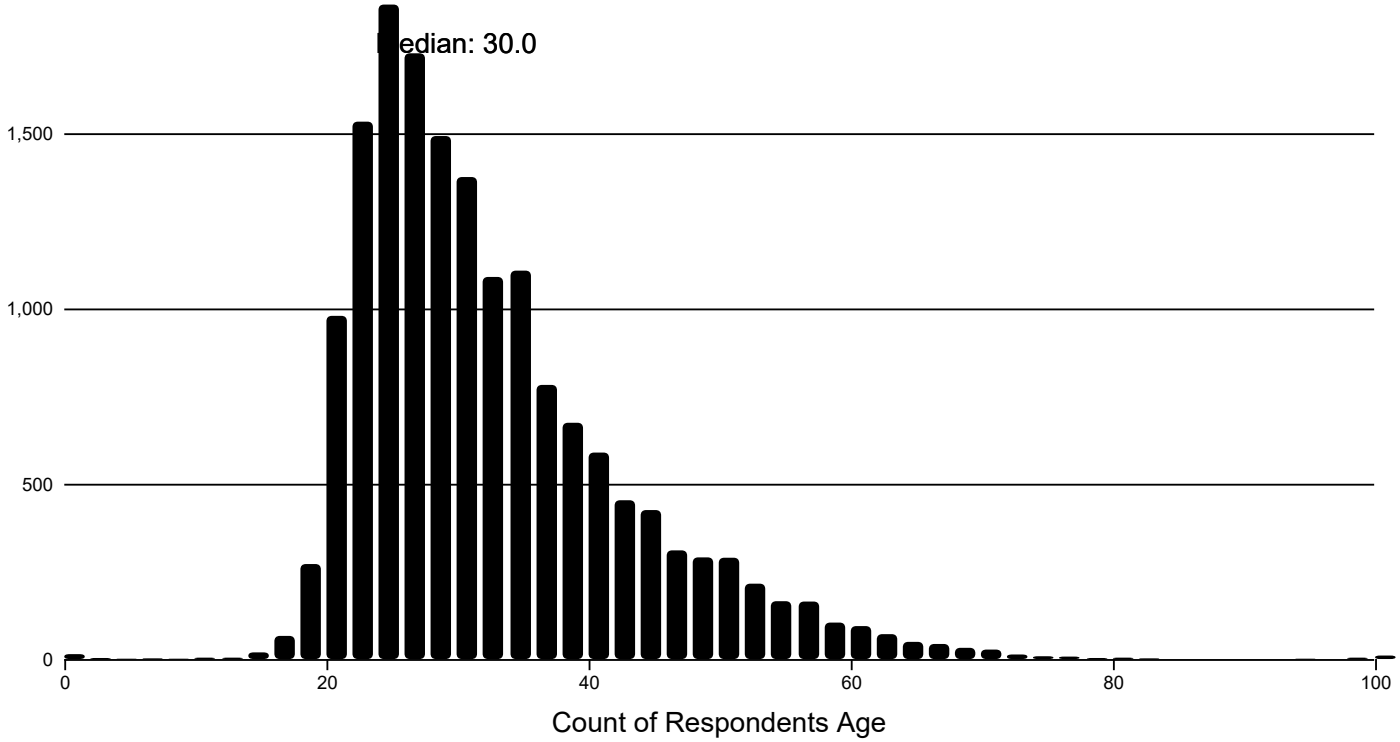
Who's working with data?

There are a lot of ways to look at who is working with data, but we'll get started with some demographic info on the jobs and backgrounds of people doing doing data science today:

How old are you?

On average, survey respondents were around 30 years old, but this value varies between countries. For instance, the average respondent from India was about 9 years younger than the average respondent from Australia .

Country ▼ Experience ▼ Job Title ▼ FILTER GENDER All Female Male Other

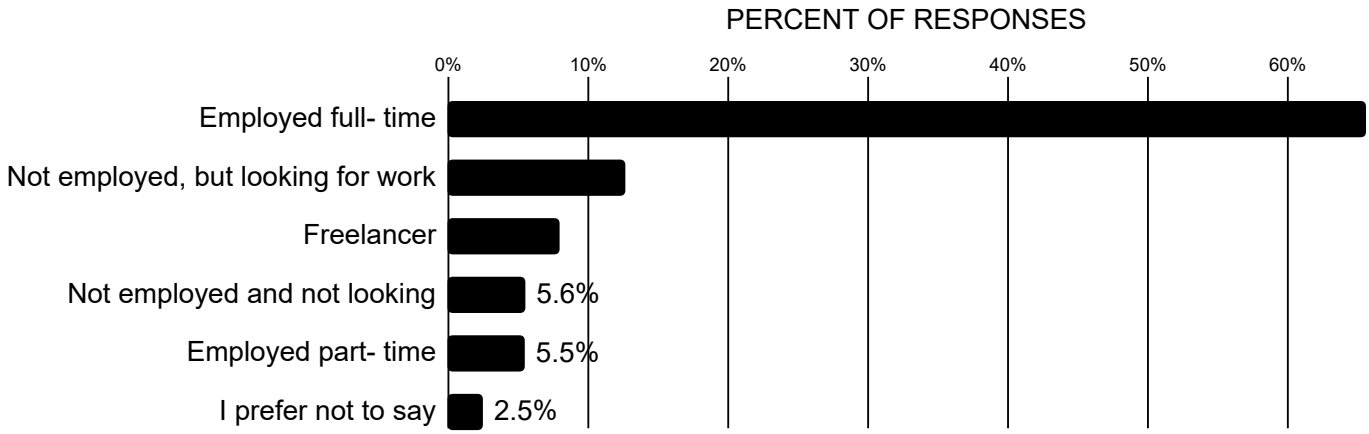


16,385 responses

View code in Kaggle Kernels

What is your employment status?

Country ▼ Job Title ▼ FILTER GENDER All Female Male Other



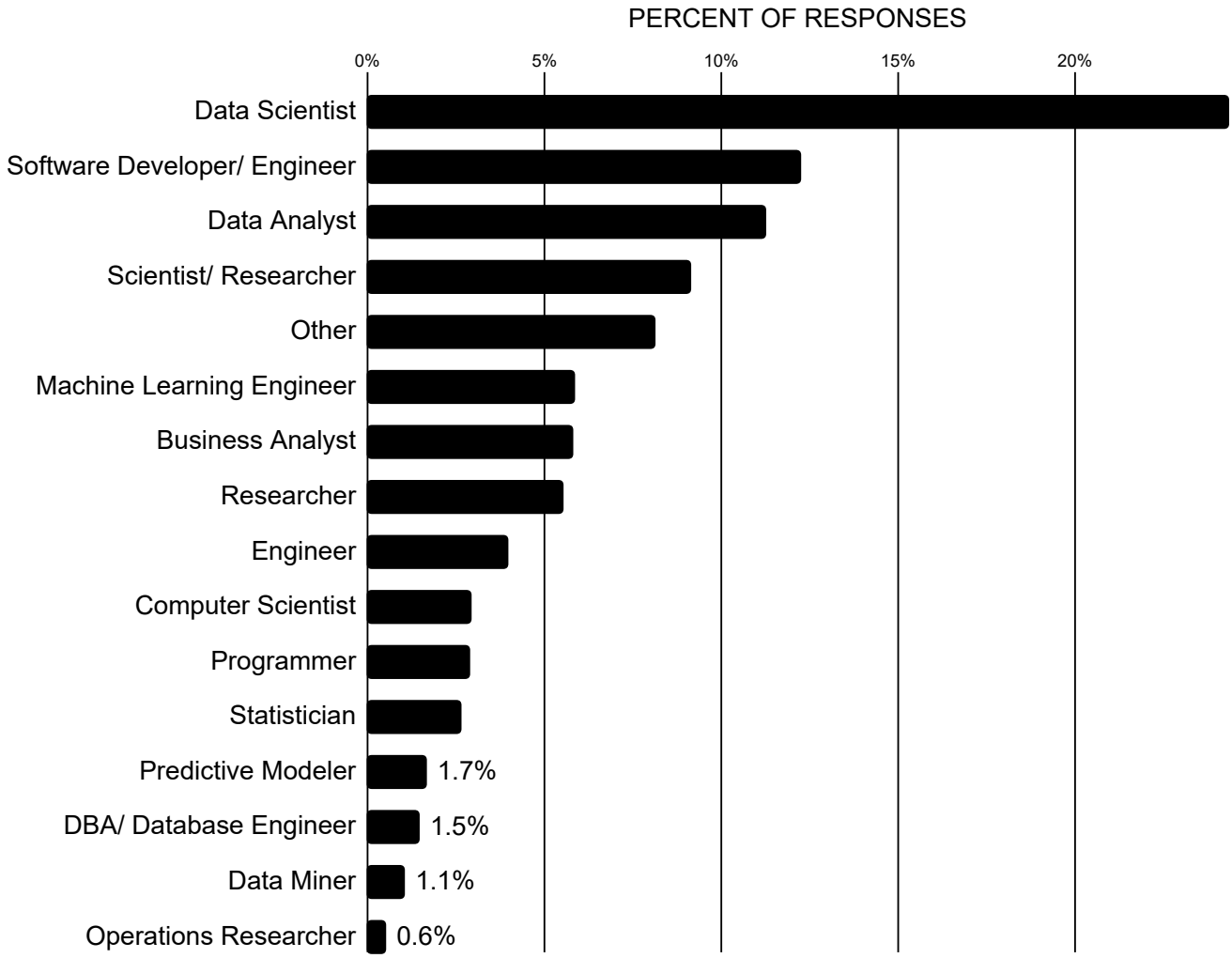
16,598 responses

View code in Kaggle Kernels

What is the title of your job?

While we define “data scientist” as someone who uses code to analyze data, we found that there are a ton of job titles that fall into the realm of data science. For example, in both Iran and Malaysia , the most popular job title for those doing data science work is “Scientist or Researcher”.

Country ▼ Experience ▼ Salary (USD) ▼ FILTER GENDER All Female Male Other



9,811 responses

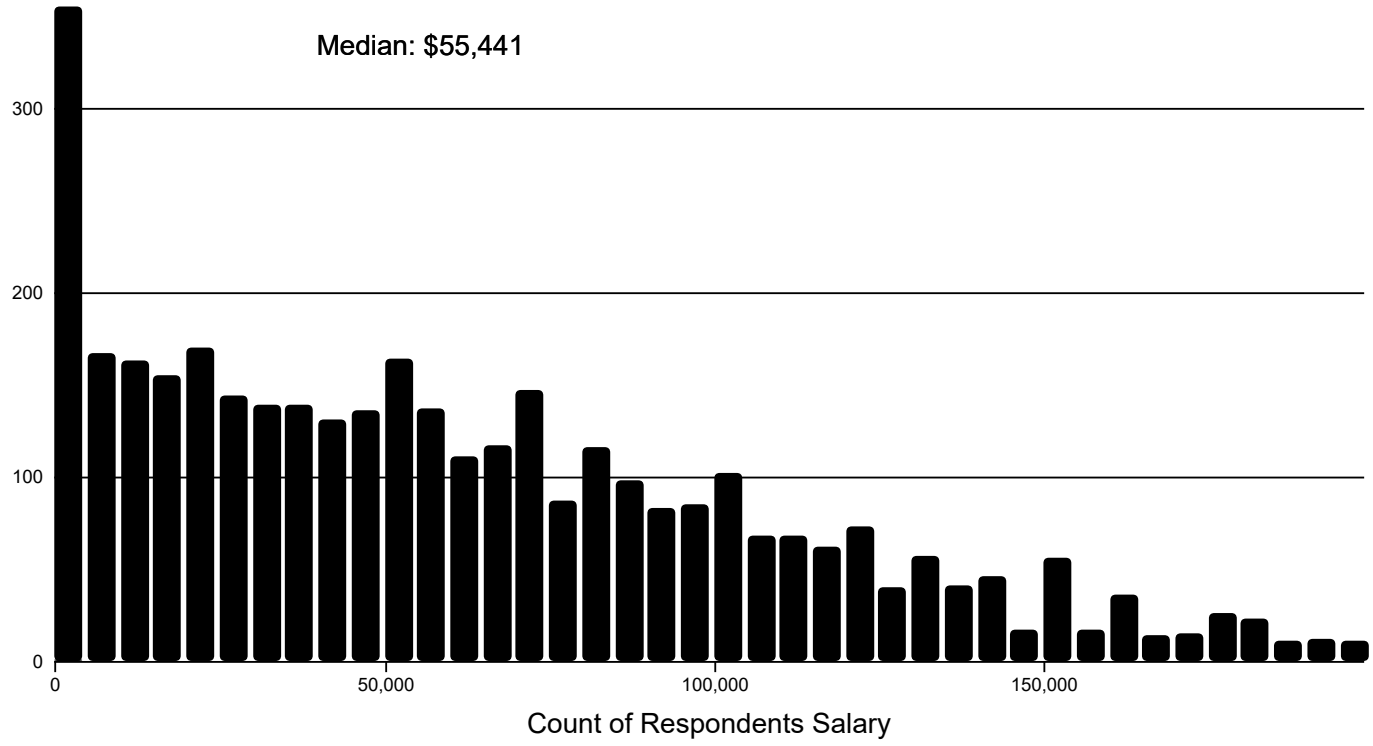
View code in Kaggle Kernels

What is your full-time annual salary?

Although “compensation and benefits” was ranked as slightly less important than “opportunities for professional development” in our survey, it’s still good to know what should be considered reasonable compensation. In the US, Machine Learning Engineers bring home the most bacon (on average).

Country ▼ Experience ▼ Job Title ▼

FILTER GENDER All Male Other Female



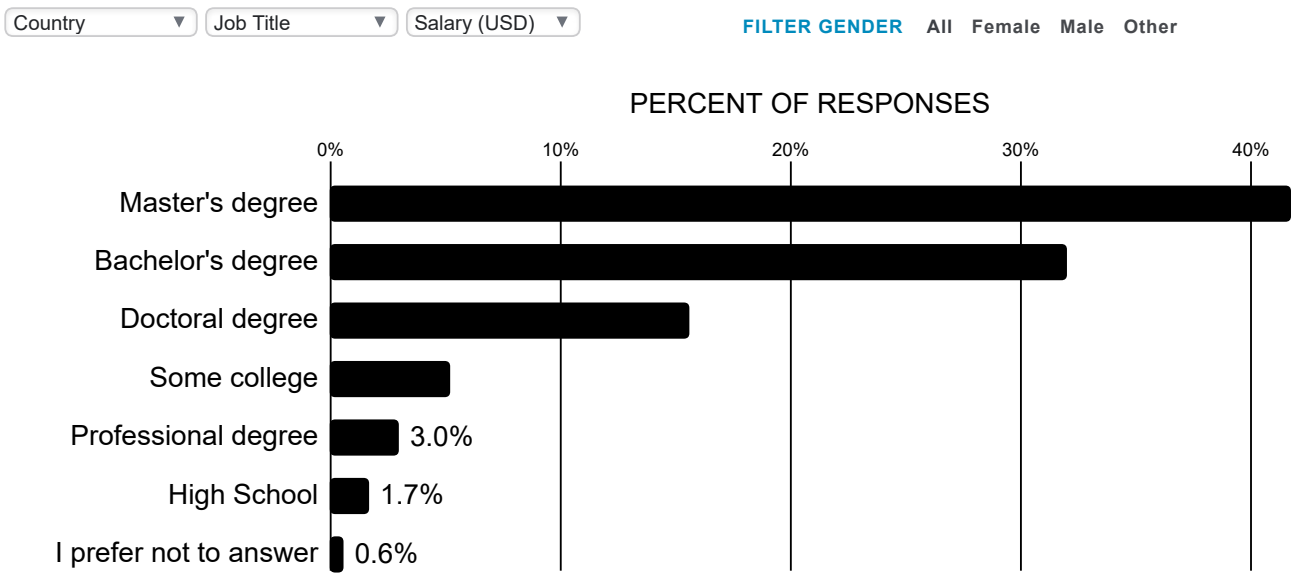
3,771 responses

131 responses are not shown that exceed the maximum value shown, but they are included in the median calculation.

View code in Kaggle Kernels

What is your highest level of formal education?

So, should you get that next degree? In general, the highest percentage of people in working data science, obtained a Master's degree. But those people in the highest salary ranges ($150K - $200K and $200k+) are just as likely to have a doctoral degree.



15,015 responses

View code in Kaggle Kernels

The average survey respondent was a 30-year-old with a Master's degree, a job as a Data Scientist, and who makes about $55,000 per year. But people are not averages. These first few demographic questions give just the surface-level view of how diverse Kaggle's data science community is in age, gender, country of residence, job title, salary, experience level, and formal education.

Share your insights, win $1,000!

There are endless ways to slice this survey data to uncover cool stories and facts. We're giving away $1,000/week in October for top analyses shared on Kaggle Kernels.

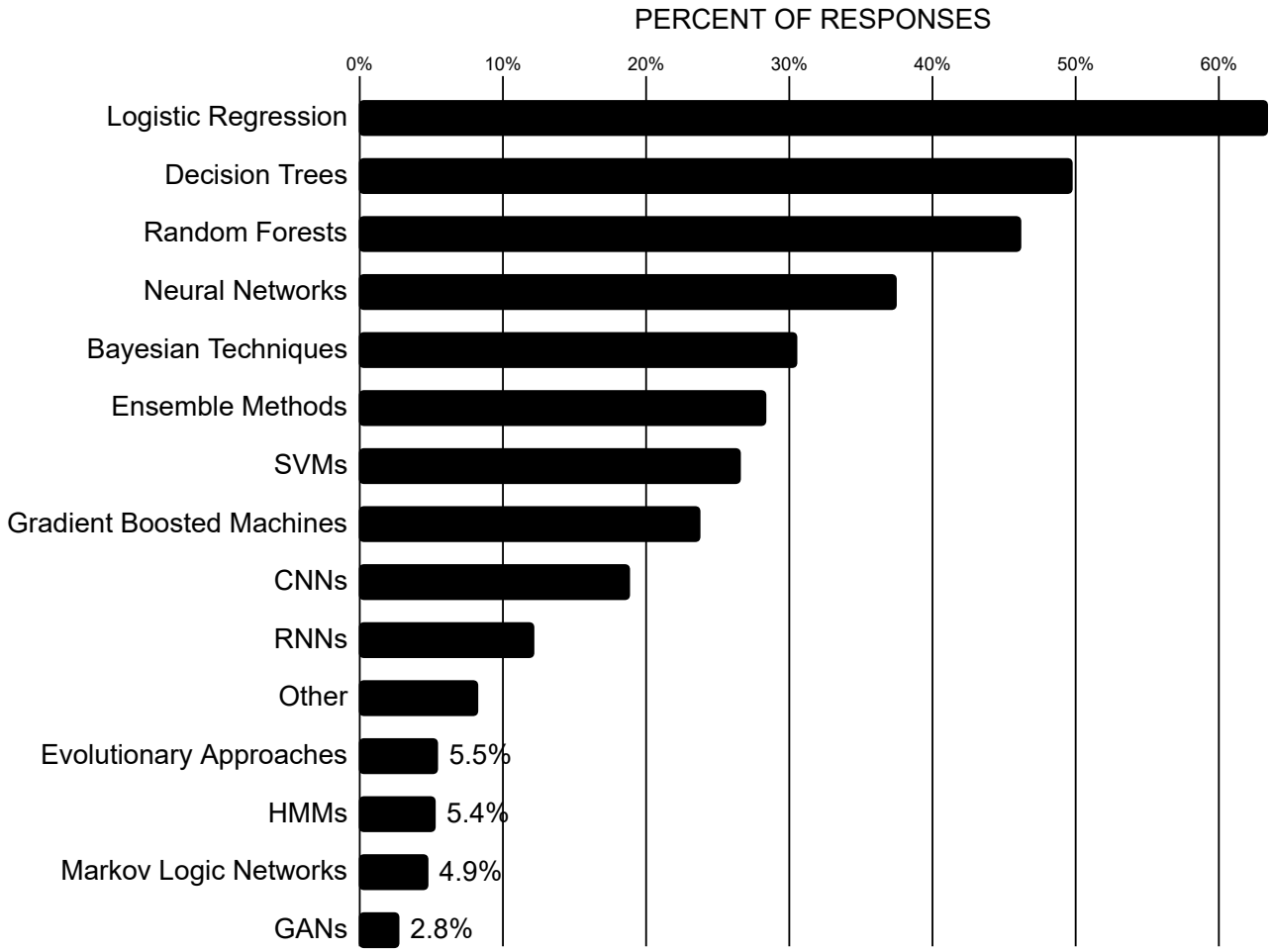
What do data scientists do at work?

We define a data scientist as someone who 'writes code to analyze data'. We asked these people what fills their day-to-day and here's some highlights of what we found out:

What data science methods are used at work?

Logistic regression is the most commonly reported data science method used at work for all industries *except* Military and Security where Neural Networks are used slightly more frequently.

Company Size ▼ Industry ▼ Job Title ▼



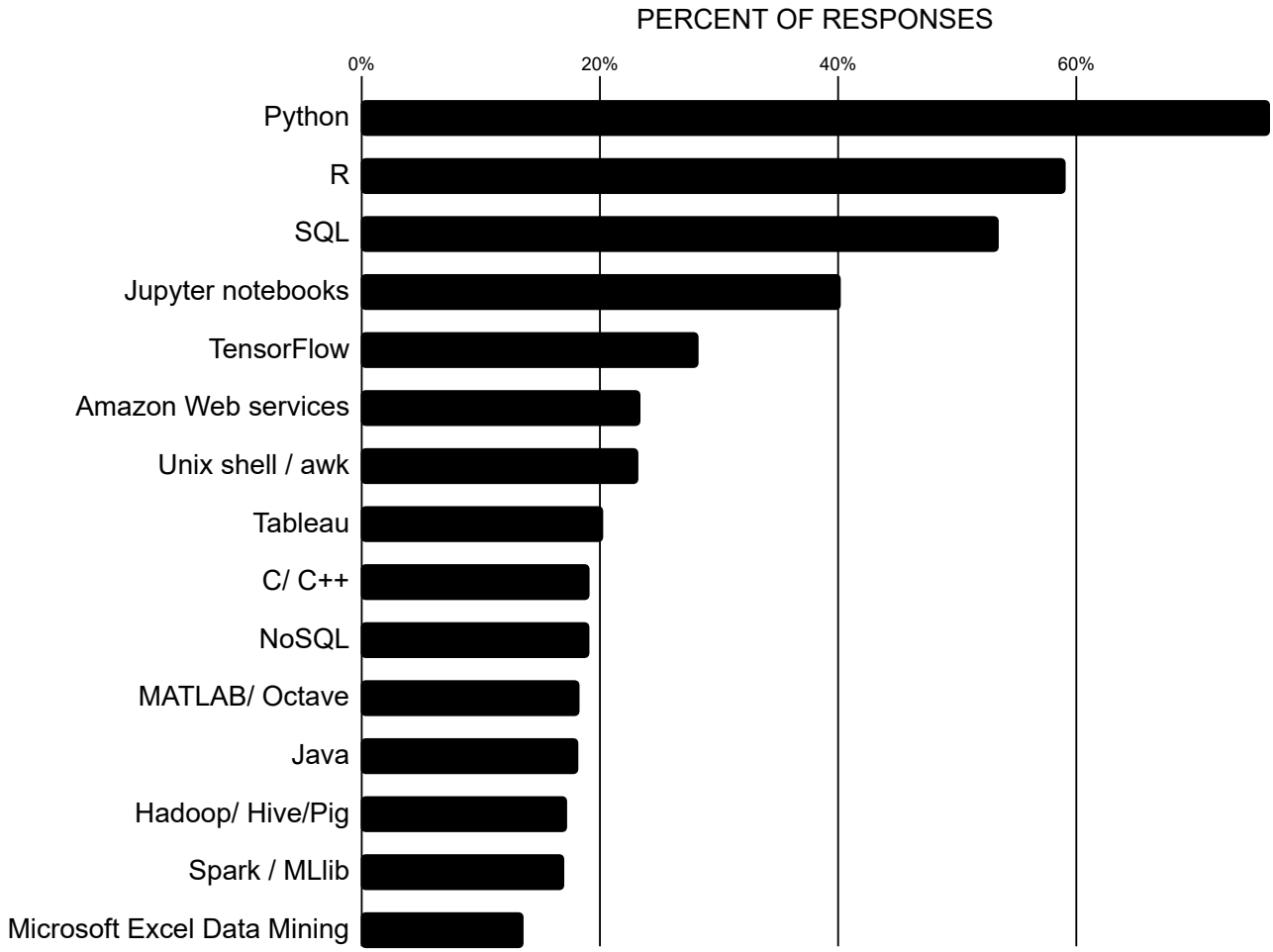
7,301 responses

View code in Kaggle Kernels

What tools are used at work?

Python was the most commonly used data analysis tool across employed data scientists overall, but more Statisticians are still loyal to R.

Company Size ▼ Industry ▼ Job Title ▼



7,955 responses

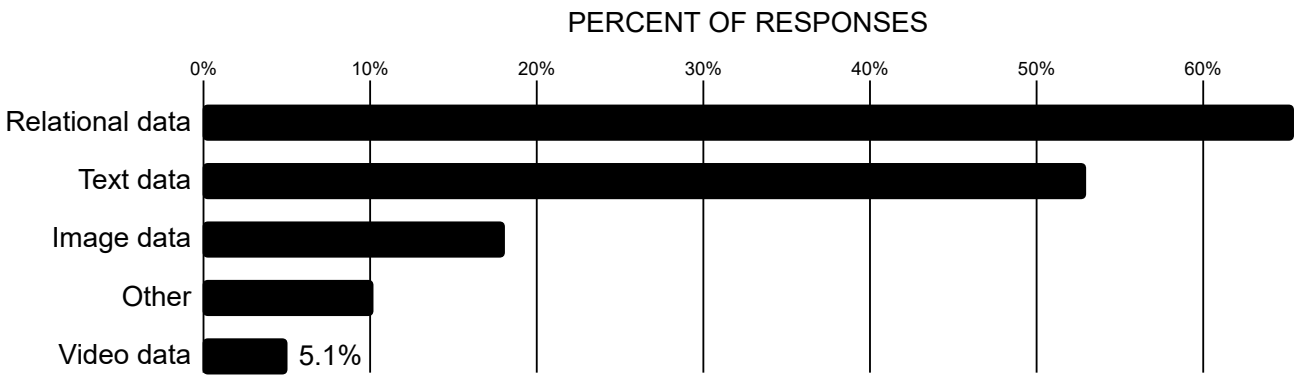
Only displaying the top 15 answers. There are 38 answers not shown.

View code in Kaggle Kernels

What type of data is used at work?

Relational data is the most commonly reported type of data used at work for all industries except for Academia and the Military and Security industry where text data's used more.

Company Size ▼ Industry ▼ Job Title ▼



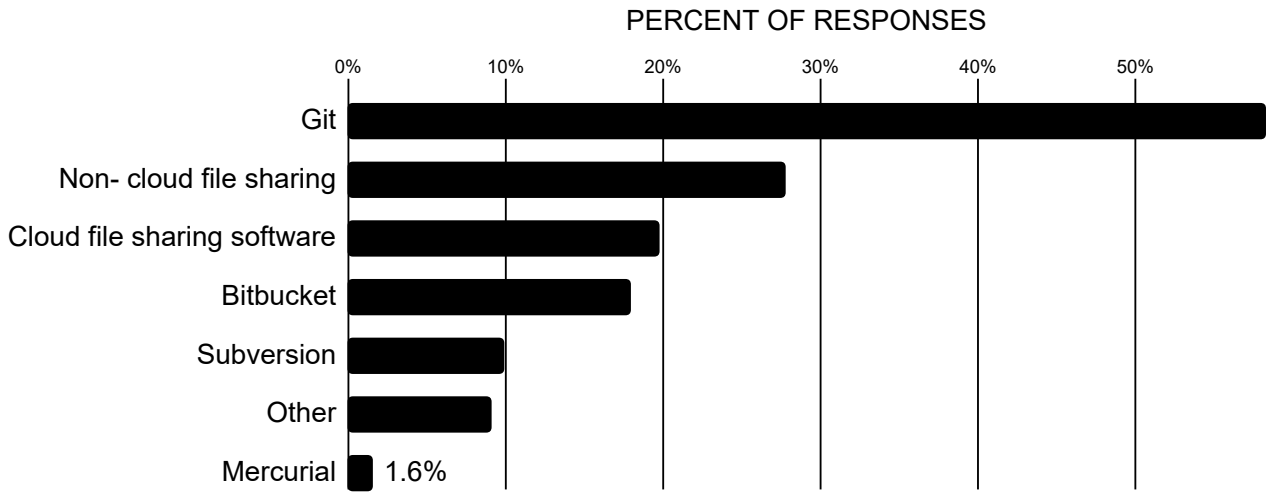
8,024 responses

View code in Kaggle Kernels

How is code shared at work?

Although the highest percentage of respondents share their code at work using Git, people in large companies are more likely to stay off the cloud and use file sharing softwares like Email. Those in the smallest companies are staying more agile by sharing in the cloud.

Company Size ▼ Industry ▼ Job Title ▼



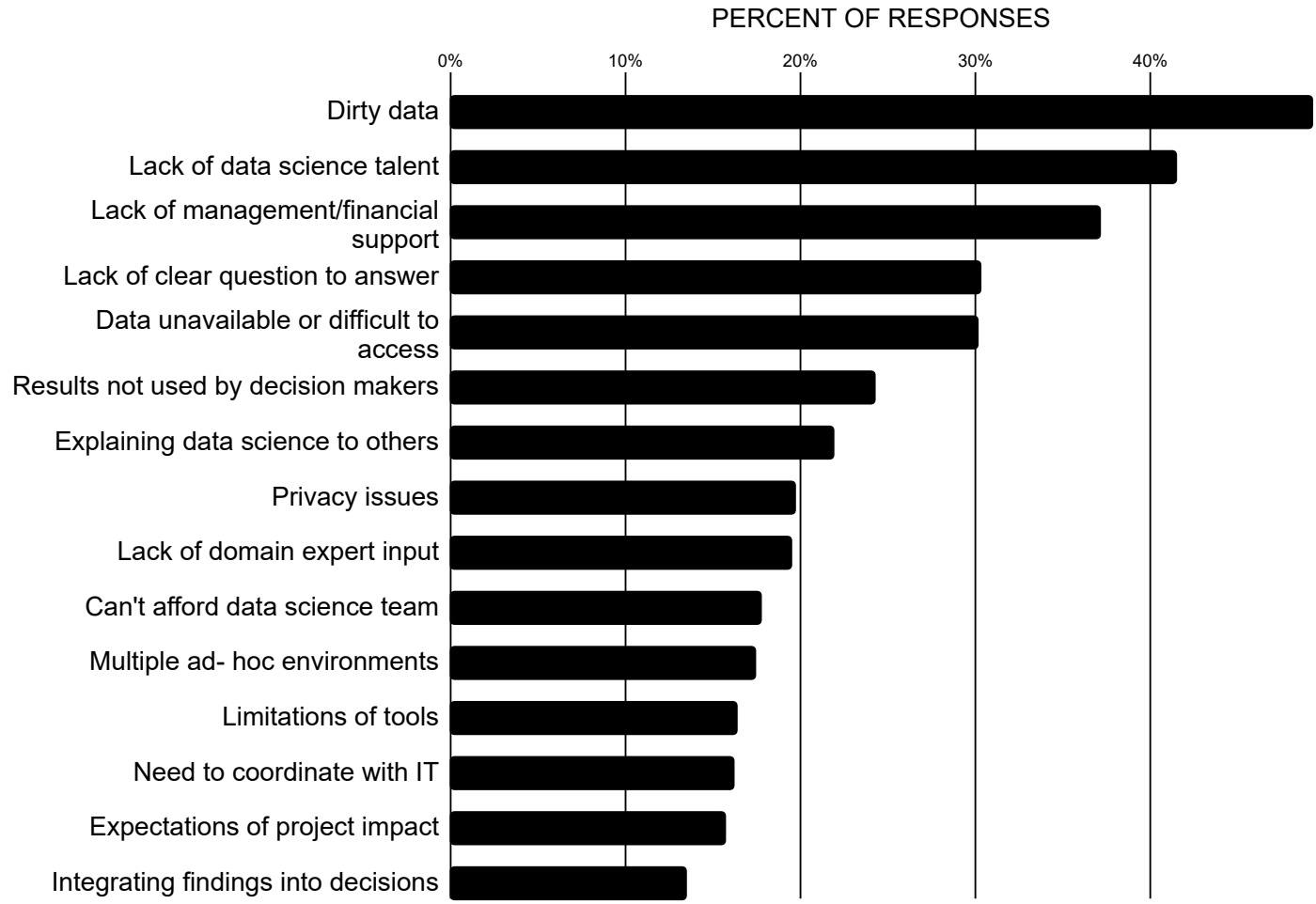
6,203 responses

View code in Kaggle Kernels

What barriers are faced at work?

Ah, dirty data, we meet again. It looks like, in general, dirty data is the most common problem for workers in the data science realm. One exception are those necessarily meticulous Database Engineers . After dirty data, company politics, lack of management and/or financial support are the real thorns in a data scientist's side.

Company Size ▼ Industry ▼ Job Title ▼



7,376 responses

Only displaying the top 15 answers. There are 7 answers not shown.

View code in Kaggle Kernels

If you clicked around the filters, you may have noticed that respondents across many industries indicated that they struggle with a lack of data science talent in their organization. That means new data scientists are in luck (if you know where to look)! Keep reading to see how our currently employed survey respondents got off on the right track:

Share your insights, win $1,000!

There are endless ways to slice this survey data to uncover cool stories and facts. We're giving away $1,000/week in October for top analyses shared on Kaggle Kernels.

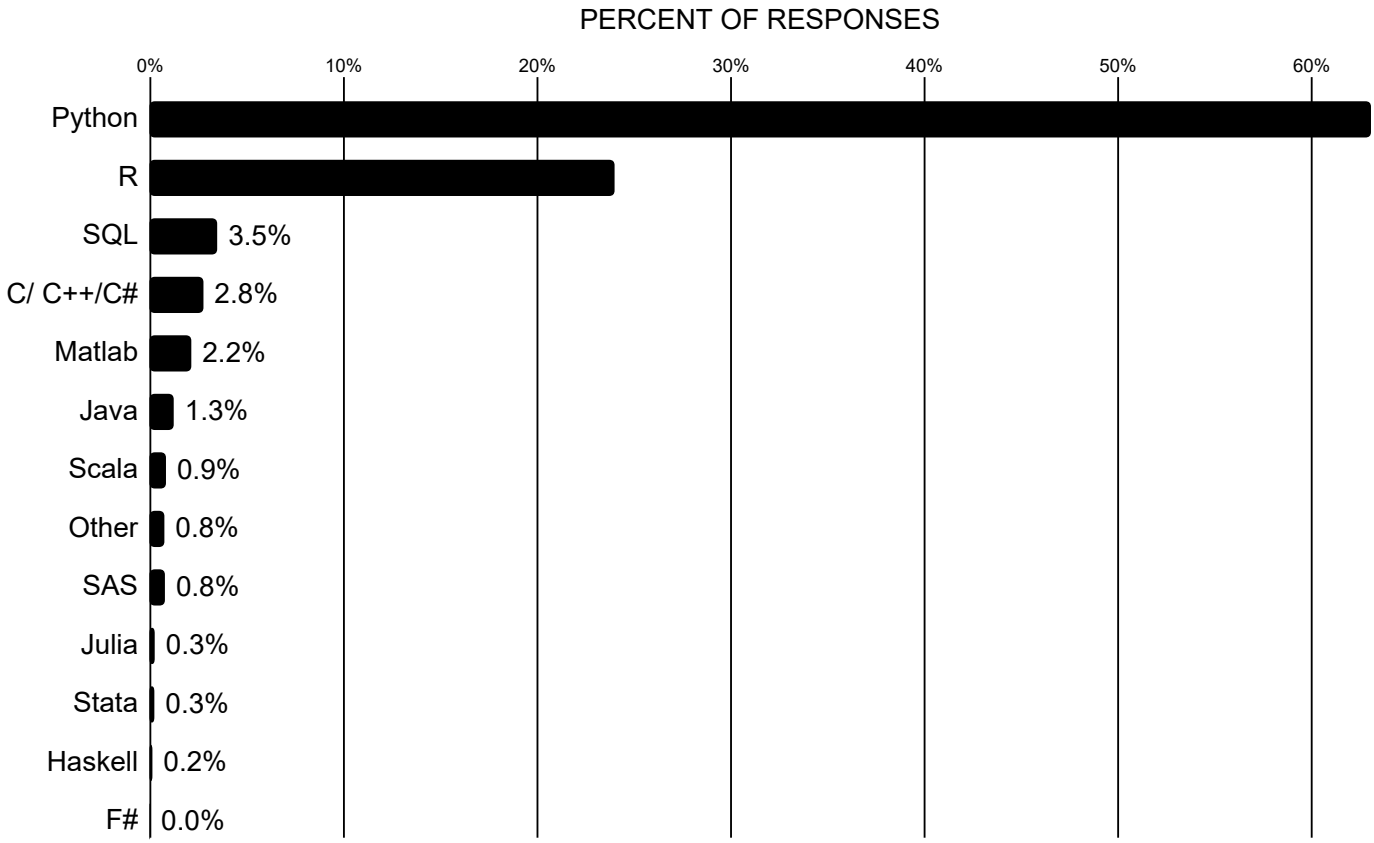
How can new data scientists break into the field?

When you're starting a new career path, it's helpful to find out how other people managed to find success. We asked people working in the data science industry how they 'made it'. Here are a few of our favorite pieces of advice:

What language would you recommend new data scientists learn first?

Everyone data scientist has an opinions on what language you should learn first. As it turns out, people who solely use Python or R feel like they made the right choice. But if you ask people that use both R and Python, they are twice as likely to recommend Python.

FILTER FROM USERS OF All Both Python R



10,998 responses

View code in Kaggle Kernels

What data science learning resources do you use?

Data science is a quickly changing field and there are a lot of valuable resources to help you learn and stay at the top of your game so you're always eminently employable. Those already working in the data science field are using Stack Overflow Q&A, Conferences, and Podcasts more frequently to stay up to date than people entering the field. If you're making content or open source software, keep in mind that people entering the field are more commonly using the Official documentation and watching YouTube videos.

EMPLOYED IN FIELD ENTERING FIELD

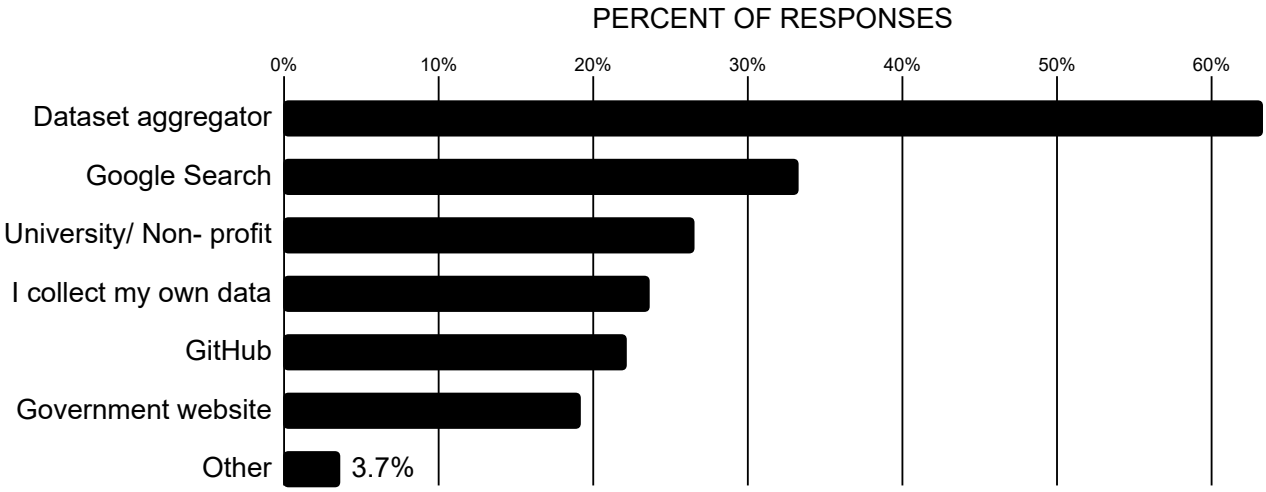
11,267 responses

View code in Kaggle Kernels

Where do you find your open data?

There's no data science without the data. When it comes to learning data science skills, knowing how to find clean open datasets to use for practice and projects is incredibly valuable. We're glad that dataset aggregators, like ours tend to be used the most frequently by members of the data science community.

FILTER EMPLOYMENT STATUS All Employed In Field Entering Field



10,796 responses

View code in Kaggle Kernels

How do you look for or find work?

When you're job hunting, it may be tempting to look for work on company websites or tech-specific job boards, but according to people who are employed in the data science realm, these are among the least helpful ways to find work. Instead, try to contact recruiters or build up your network to break into the field.

EMPLOYED IN FIELD ENTERING FIELD

11,467 responses

View code in Kaggle Kernels

Share your insights, win $1,000!

There are endless ways to slice this survey data to uncover cool stories and facts. We're giving away $1,000/week in October for top analyses shared on Kaggle Kernels.

Kaggle is a platform for data scientists to connect, learn, find and explore data, and compete in machine learning challenges. Since our launch in 2010, Kaggle's platform has attracted a diverse set of data scientists and machine learning engineers. Since then, we've registered more than one million users from almost every corner of the earth and become the world's largest data science community.

All data was obtained through the Kaggle 2017 Data Science Survey. The cleaned data can be accessed here and a kernel with the code we used to analyze the data can be found here.

Countries that had fewer than 50 respondents were grouped together into a single "Other" category. Some responses have been shortened for the purposes of these visualizations, but the entire responses and original questions can be found here. Some questions allowed respondents to select multiple answers. Bar charts displaying these questions may add up to more than 100%.